# Tema 2.2

# TEORÍA DE LA ESTIMACIÓN

Febrero-Mayo 2006



# INDICE

# 2.2.1 INTRODUCCIÓN

La clasificación bayesiana precisa del conocimiento de $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i)$ y de $\Pr(\omega_i)$. Para el cálculo de estas magnitudes se requiere:

- Disponer de una serie de datos previamente clasificados de forma fiable.
- Disponer de un estimador de esas probabilidades.

La estimación de $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i)$ requiere muchos datos a menos que podamos definir una función que dependa de unos pocos parámetros $\boldsymbol{\theta}_i$.

Caso gaussiano: $\boldsymbol{\theta}_i$ contiene la media y la matriz de covarianza

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\omega_i,\boldsymbol{\theta}_i)=\frac{1}{(2\pi)^{d/2}\,|\mathbf{C}_i|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}$$

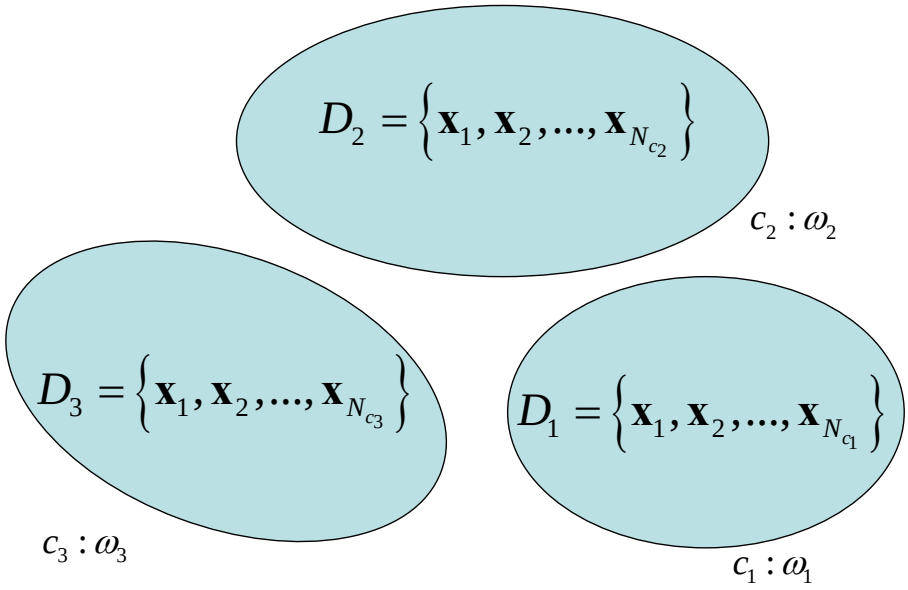


Existen dos alternativas:

1. Estimación de máxima verosimilidud (ML): Los parámetros a estimar se consideran deterministas (aunque desconocidos).

2. Estimación bayesiana: Los parámetros son variables aleatorias con distribución a priori conocida. La definición del estimador bayesiano permite mejorar fácilmente la estimación de $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$ cuando se dispone de nuevos datos.

Supondremos que disponemos de una base de datos clasificada (un conjunto de vectores de características clasificados por categorías), a partir de las cuales hemos de determinar $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$:






# 2.2.2 ESTIMACIÓN DE MÁXIMA VEROSIMILITUD (ML)

Si los datos observados en cada clase *D* son independientes:

$$f(D_i \mid \boldsymbol{\theta}_i) = \prod_{k=1}^{N_{c_i}} f_{\mathbf{x}}(\mathbf{x}_k \mid \boldsymbol{\theta}_i)$$

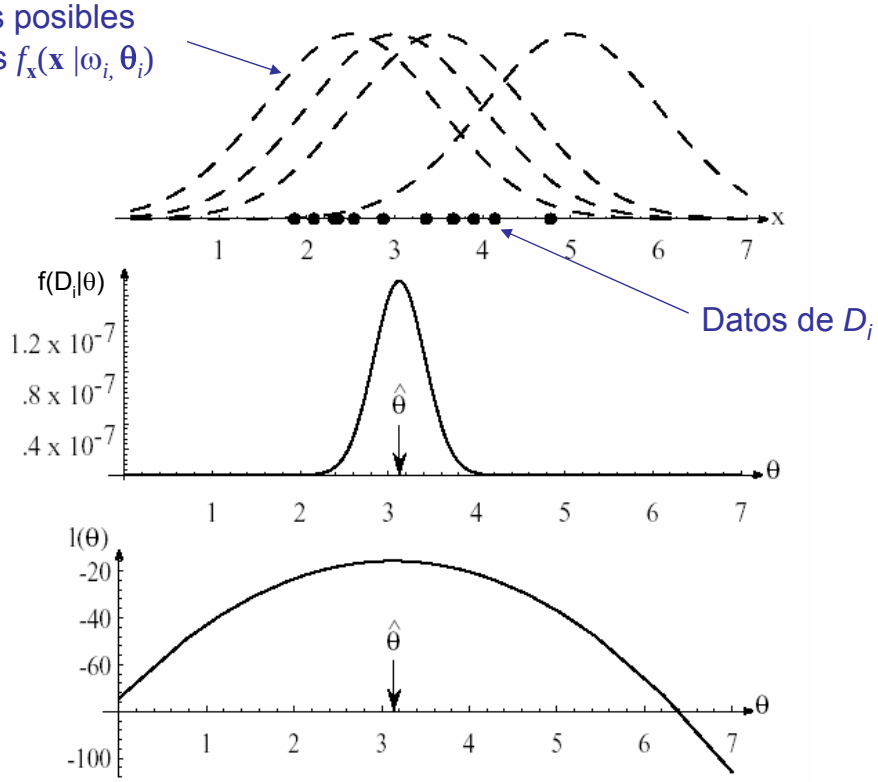
es la función de verosimilitud. El estimador maximiza esta función (o su logaritmo):

$$\hat{\boldsymbol{\theta}}_{i,ML} = \arg\max_{\boldsymbol{\theta}_i} f(D_i \mid \boldsymbol{\theta}_i) = \arg\max_{\boldsymbol{\theta}_i} \ln f(D_i \mid \boldsymbol{\theta}_i)$$

Un conjunto de condiciones necesarias (no suficientes) para obtener el estimador vienen dadas por:

$$\nabla_{\boldsymbol{\theta}_i} \ln f(D_i \mid \boldsymbol{\theta}_i) = \mathbf{0}$$

Algunas posibles funciones $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i, \boldsymbol{\theta}_i)$




Datos de $D_i$



## Propiedades del estimador ML:

1. Es asintóticamente insesgado (en muchos casos es insesgado aunque N sea pequeño).
2. Es asintóticamente eficiente (cuando N es grande, su varianza es la de Crámer-Rao).

## Sin embargo

1. No tiene porqué ser el que proporcione menor error de clasificación cuando utilicemos

$$f_{\mathbf{x}}(x \,|\, \omega_i, \hat{\boldsymbol{\theta}}_{i,ML})$$

2. Si la pdf asumida es muy distinta de la real las estimaciones pueden ser de poca calidad.

## Ejemplo 1:

Estimador ML de la media $\boldsymbol{\mu}_i$ si la matriz de covarianza $\mathbf{C}_i$ es conocida, en el caso gausiano multivariable. Demostrad que:

$$\hat{\boldsymbol{\mu}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \mathbf{x}_k$$

## Ejemplo 2:

Estimador ML de la media $\boldsymbol{\mu}_i$ y la matriz de covarianza $\mathbf{C}_i$ en el caso gausiano multivariable. Demostrad que:

$$\hat{\boldsymbol{\mu}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \mathbf{x}_k \qquad \hat{\mathbf{C}}_{i,ML} = \frac{1}{N_{c_i}} \sum_{k=1}^{N_{c_i}} \left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{i,ML}\right)\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{i,ML}\right)^T$$



## Ejemplo 3:

Estimador ML de la probabilidad $p_k$ de aparición de '1' para cada una de las componentes del vector de datos binarios $\mathbf{x} \in \{0,1\}^d$ :

$$f_{\mathbf{x}}(D \mid \omega, \mathbf{p}) = \prod_{j=1}^{N_i} \prod_{k=1}^{d} p_k^{\,x_{k,j}} (1 - p_k)^{1 - x_{k,j}}$$

$$\mathbf{p} = \left[ p_1, \ldots, p_d \right]$$

# 2.2.3 ESTIMACIÓN BAYESIANA

Si se dispone de algún conocimiento a priori de los rangos de valores más probables de $\theta_i$ podemos aprovecharlo para:

1. Mejorar la estimación ML de $\theta_i$ (usando MAP)

$$\hat{\boldsymbol{\theta}}_{i,MAP} = \arg\max_{\boldsymbol{\theta}_i} f\left(D_i \mid \boldsymbol{\theta}_i\right) f\left(\boldsymbol{\theta}_i\right) = \arg\max_{\boldsymbol{\theta}_i} \left[\ln f\left(D_i \mid \boldsymbol{\theta}_i\right) + \ln f\left(\boldsymbol{\theta}_i\right)\right]$$

2. Estimar directamente las probabilidades a posteriori $\Pr(\omega_i|\mathbf{x})$

Calculando $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$ y $\Pr(\omega_i)$. Es el procedimiento más aconsejable en una aplicación de clasificación.



## Suposiciones

Queremos determinar la probabilidad a posteriori a partir de las observaciones en $D_i$, y supondremos que:

- La forma de $f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}_i)$ es conocida pero no el parámetro $\theta_i$
- Nuestro conocimiento a priori de $\theta_i$ está en $f(\theta_i)$
- El resto de nuestro conocimiento sobre $\theta_i$ viene dado por los datos en $D_i$

## Procedimiento:

**1**. Promediar la forma conocida para la función de verosimilitud respecto a la probabilidad a posteriori del parámetro:

$$f_{\mathbf{x}}(\mathbf{x} \mid \omega_i) \cong f_{\mathbf{x}}(\mathbf{x} \mid D_i) = \int f(\mathbf{x} \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i \mid D_i) d\boldsymbol{\theta}_i$$

**2**. Calculamos la probabilidad a posteriori del parámetro como

$$f(\boldsymbol{\theta}_i \mid D_i) = \frac{f(D_i \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)}{\int f(D_i \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i} \propto f(D_i \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)$$

**3**. Suponiendo independencia de los datos en $D_i$

$$f(D_i \mid \boldsymbol{\theta}_i) = \prod_{k=1}^{N_i} f(\mathbf{x}_k \mid \boldsymbol{\theta}_i)$$



## Ejemplo 4:

Estimador bayesiano de $f_{\mathbf{x}}(\mathbf{x}|D_i)$ si

$$f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \mathbf{C}) \qquad f(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \mathbf{C}_0)$$

donde se suponen conocidas $\boldsymbol{\mu}_0$, $\mathbf{C}_0$ y $\mathbf{C}$, y se dispone de los datos observados $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \omega$

A partir de **2** y **3** podemos escribir:

$$f(\boldsymbol{\mu}|D) = \alpha \prod_{k=1}^{N} f_{\mathbf{x}}(\mathbf{x}_k \mid \boldsymbol{\mu}) f(\boldsymbol{\mu}) =$$

$$= \alpha' \exp\left[ -\frac{1}{2} \left( \boldsymbol{\mu}^T \left( N\mathbf{C}^{-1} + \mathbf{C}_0^{-1} \right) \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \left( \mathbf{C}^{-1} \sum_{k=1}^{N} x_k + \mathbf{C}_0^{-1} \boldsymbol{\mu}_0 \right) \right) \right]$$

La ecuación puede escribirse también como:

$$f(\boldsymbol{\mu}|D) = \alpha'' \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T \mathbf{C}_N^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)\right]$$

donde:

$$\mathbf{C}_N^{-1} = N\mathbf{C}^{-1} + \mathbf{C}_0^{-1}$$
$$\mathbf{C}_N^{-1}\boldsymbol{\mu}_N = N\mathbf{C}^{-1}\mathbf{m}_N + \mathbf{C}_0^{-1}\boldsymbol{\mu}_0$$
$$\mathbf{m}_N = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

Aplicando la igualdad matricial:

$$\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1} = \mathbf{A}\left(\mathbf{A}+\mathbf{B}\right)^{-1}\mathbf{B} = \mathbf{B}\left(\mathbf{A}+\mathbf{B}\right)^{-1}\mathbf{A}$$



pueden reformularse la media y la matriz de covarianza:

$$\boldsymbol{\mu}_N = \mathbf{C}_0\left(\mathbf{C}_0+\frac{1}{N}\mathbf{C}\right)^{-1}\mathbf{m}_N + \frac{1}{N}\mathbf{C}\left(\mathbf{C}_0+\frac{1}{N}\mathbf{C}\right)^{-1}\boldsymbol{\mu}_0$$
$$\mathbf{C}_N = \mathbf{C}_0\left(\mathbf{C}_0+\frac{1}{N}\mathbf{C}\right)^{-1}\frac{1}{N}\mathbf{C}$$

Nótese que la media es una combinación lineal del conocimiento a priori de la media $\boldsymbol{\mu}_0$ y la información aportada por los datos $\mathbf{m}_N$. Integrando la ecuación **1**:

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\omega) \cong f_{\mathbf{x}}(\mathbf{x}\,|\,D) = \int f(\mathbf{x}\,|\,\boldsymbol{\mu})\, f(\boldsymbol{\mu}\,|\,D)\, d\boldsymbol{\mu} \sim N\left(\boldsymbol{\mu}_N, \mathbf{C}+\mathbf{C}_N\right)$$

Cuando $N \to \infty$ la estimacion tiende a ser ML

$$\boldsymbol{\mu}_N = \mathbf{m}_N \qquad\qquad \mathbf{C}_N = \frac{1}{N}\mathbf{C}$$

## Comparación:

La función $f(D_i \mid \boldsymbol{\theta}_i)$ tendrá un pico tanto más abrupto alrededor de $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ cuanto mayor sea $N_i$.

Si $f(\boldsymbol{\theta}_i)$ no es cero y no varía mucho cerca de $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ entonces

$$f(\boldsymbol{\theta}_i \mid D_i) = \frac{f(D_i \mid \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i)}{f(D_i)}$$

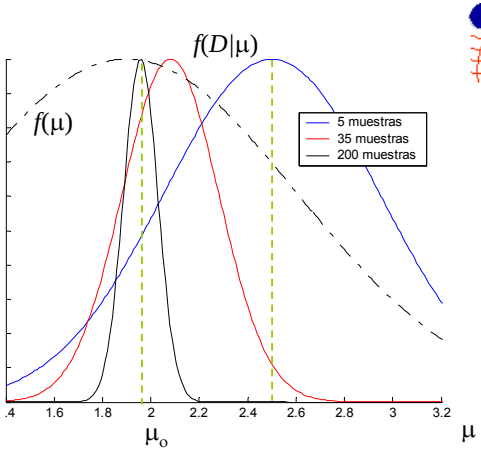
tambien tiene un pico en $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ y los estimadores obtenidos por Bayes y mediante ML coinciden. En la práctica, si el número de muestras de entrenamiento es pequeño, es mejor la estimación Bayesiana.



## Ejemplo 5:

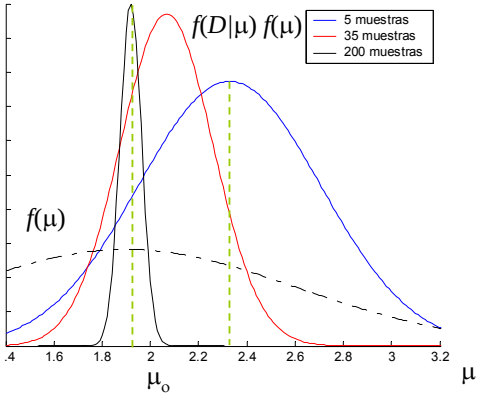
Estimación **ML** de la media ($\mu_0$=2) sobre un número variable de muestras Gaussianas.

La fdp a priori de $\mu$ es Gaussiana.



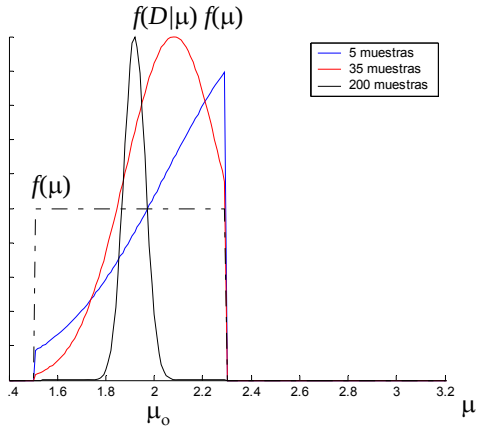

Estimación **Bayesiana** de la media ($\mu_0$=2) sobre un número variable de muestras Gaussianas.

La fdp a priori de $\mu$ es Gaussiana.



Estimación **Bayesiana** de la media ($\mu_0$=2) sobre un número variable de muestras Gaussianas.

La fdp a priori de $\mu$ es uniforme.






# 2.2.4 PROBLEMAS DE DIMENSIONALIDAD

Al aumentar el número de características independientes en un problema de clasificación es posible hacer tender el error a cero.

Ejemplo 1:

Problema de clasificación con dos clases gaussianas equiprobables

$$f_{\mathbf{x}}(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i; \boldsymbol{\Sigma}) \qquad i = 1, 2$$
$$\Pr(\omega_1) = \Pr(\omega_2)$$

$$\mathbf{x} \in \mathbb{R}^d$$

La probabilidad de error queda definida como:

$$\Pr(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} \exp(-u^2/2)du$$

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

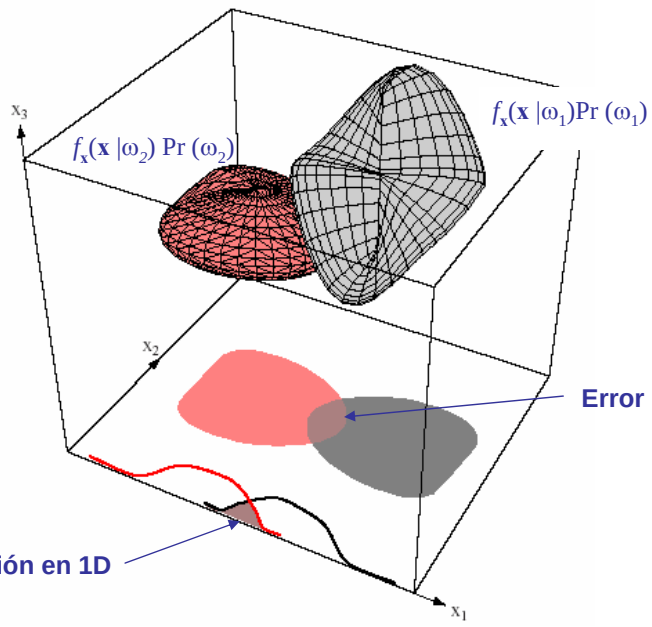
Si las características son independientes entre sí:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix} \qquad r^2 = \sum_{i=1}^{d} \left( \frac{\boldsymbol{\mu}_{i,1} - \boldsymbol{\mu}_{i,2}}{\sigma_i} \right)^2$$

Cuando d→∞, r→∞, $\Pr(e) \rightarrow 0$



Ejemplo 2:




En este caso, las pdf de dos clases están solapadas cuando se dispone de una ($x_1$) o dos características ($x_1$, $x_2$). Cuando se añade la tercera característica ($x_3$), las dos pdf aparecen completamente separadas y el error de clasificación se reduce a cero.

⇒ Sin embargo, disponer de características adicionales no siempre mejora la clasificación si se dispone de pocas observaciones para determinar parámetros adicionales.

⇒ Por otra parte, es posible que algunas de las características:

- No aporten información que permita clasificar mejor
- No sean independientes de otras características
- No se ajusten a la pdf asumida

⇒ Disponer de muchas características aumenta la complejidad de la solución.

**La definición del número de características relevantes es necesaria en todo sistema de clasificación**



Reducción del número de características mediante combinación lineal:

$$\mathbf{y}_k = \mathbf{W}^T \mathbf{x}_k \qquad \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^{d'}, \mathbf{W}^T \in \mathbb{R}^{d' \times d} \qquad d' < d$$

Soluciones posibles para $\mathbf{W}$:

1. Proyectar los vectores $\mathbf{x}_k$ de la mejor forma posible en el sentido del error cuadrático
   ⇒ análisis de componentes principales (PCA)

2. Proyectar los vectores $\mathbf{x}_k$ de forma que las clases resultantes queden lo más separadas posible
   ⇒ análisis por múltiples discriminantes (MDA)

# ANÁLISIS DE COMPONENTES PRINCIPALES

Disponemos de un total de *N* observaciones asociadas a todas las clases:

$$D = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\} \qquad \mathbf{x} \in \mathbb{R}^d$$

buscamos una matriz $\mathbf{W}$ unitaria ($\mathbf{W}^T\mathbf{W} = \mathbf{I}$) que mejor aproxime

$$\mathbf{x}_k \cong \mathbf{W}\mathbf{y}_k + \mathbf{m} \qquad \mathbf{m} = \frac{1}{N}\sum_{\mathbf{x}\in D}\mathbf{x}$$

en el sentido del error cuadrático medio

$$J = \sum_{k=1}^{N} \left\| \mathbf{W}\mathbf{y}_k + \mathbf{m} - \mathbf{x}_k \right\|_2^2 \qquad (1)$$



Los vectores de dimensión reducida vienen dados por:

$$\mathbf{y}_k = \mathbf{W}^T\left(\mathbf{x}_k - \mathbf{m}\right) \qquad (2)$$

y la matriz de transformación es $\mathbf{W} = \left[\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_{d'}\right]$

donde los vectores $\mathbf{w}_j$ cumplen:

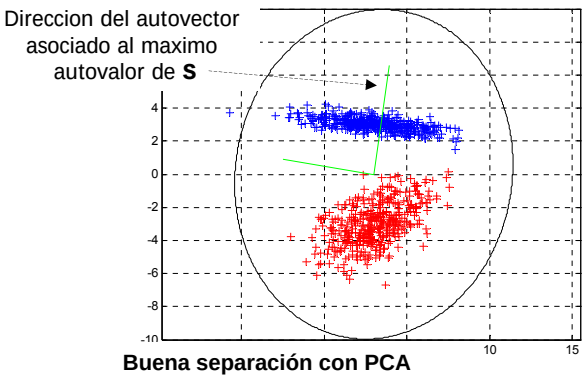
$$\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$$

$$\mathbf{S} = \sum_{\mathbf{x}\in D_i}\left(\mathbf{x}-\mathbf{m}\right)\left(\mathbf{x}-\mathbf{m}\right)^T$$

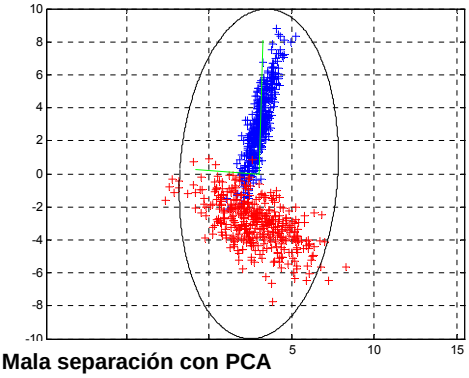
La minimización del error cuadrático medio implica utilizar en $\mathbf{W}$ los autovectores asociados a los mayores autovalores.

**** Demostradlo, minimizando (1) con la restricción $\mathbf{w}_j^T\mathbf{w}_j = 1$ ****

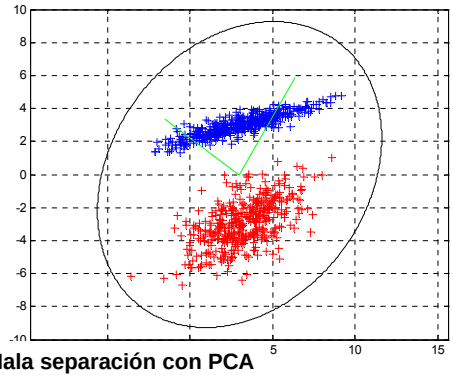
Los autovectores definen las direcciones principales de un hiperelipsoide. Son ortogonales y apuntan en las direcciones de máxima dispersión.




**Buena separación con PCA**



**Mala separación con PCA**



**Mala separación con PCA**

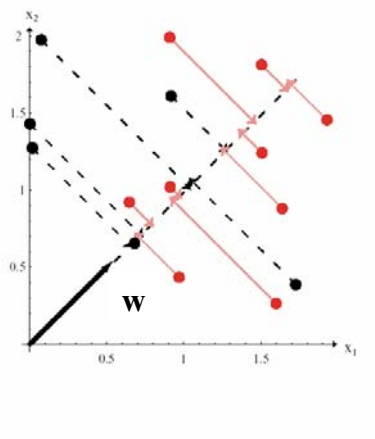


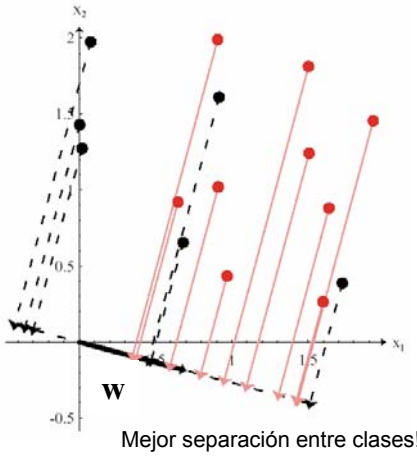
# ANÁLISIS POR MÚLTIPLES DISCRIMINANTES

La transformación por componentes principales no es siempre útil para discriminar entre distintas clases. Sería mejor definir una transformación que

- Aumente la distancia inter-clases y
- Disminuya la dispersión intra-clase.

Ejemplo 1:




Mejor separación entre clases!!

Construiremos una transformación $\mathbf{W}^T$ desde un espacio de dimension *d* (tamaño de los vectores observados $\mathbf{x}$) a un espacio de dimension $d' \leq c-1$ (donde *c* es el numero de clases).

Necesitamos una medida de la distancia inter-clases y una medida de la dispersión intra-clase, para lo cual definimos:

$$\mathbf{m}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$$ Media de los datos de la clase *i*

$$\mathbf{m} = \frac{1}{N}\sum_{\mathbf{x}\in\{D_1,\ldots,D_c\}}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{c} N_i\mathbf{m}_i$$ Media de todos los datos

$$\mathbf{S}_T = \sum_{\mathbf{x}\in\{D_1,\ldots,D_c\}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T$$ Dispersión total de los datos



$$\mathbf{S}_T = \sum_{\mathbf{x}\in\{D_1,\ldots,D_c\}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T =$$

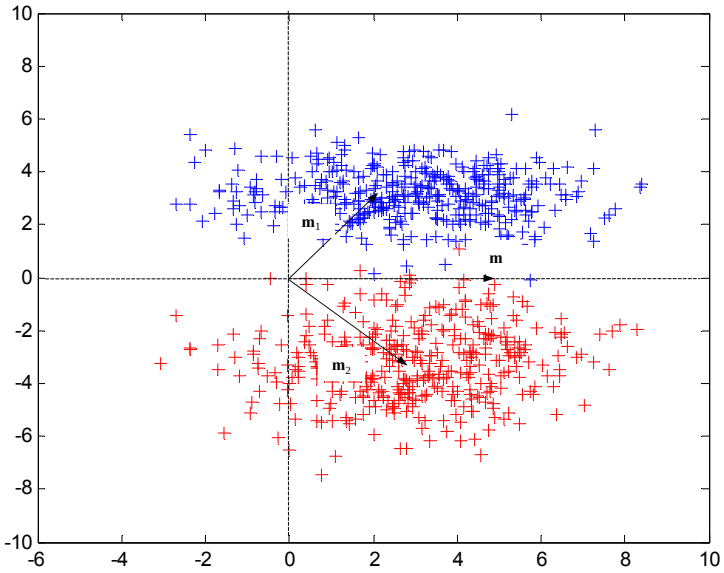
$$= \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^T =$$

$$= \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T + \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T =$$

$$= \sum_{i=1}^{c}\mathbf{S}_{C,i} + \sum_{i=1}^{c} N_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T = \mathbf{S}_C + \mathbf{S}_B$$

**Suma de matrices de dispersión intra-clases**

**Matriz de distancia inter-clases**






La transformación a aplicar será:

$$\mathbf{y}_k = \mathbf{W}^T \mathbf{x}_k \qquad \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^{d'}, \mathbf{W}^T \in \mathbb{R}^{d' \times d} \qquad d' < d$$

Las matrices de dispersión intra-clase e inter-clases quedan modificadas:

$$\mathbf{S}_C \rightarrow \mathbf{W}^T \mathbf{S}_C \mathbf{W}$$

$$\mathbf{S}_B \rightarrow \mathbf{W}^T \mathbf{S}_B \mathbf{W}$$

Si suponemos que $\mathbf{y}$ es gaussiano, el volumen de las pdf asociadas a cada clase es proporcional a $\left|\mathbf{W}^T \mathbf{S}_C \mathbf{W}\right|^{1/2}$ y la distancia entre clases es proporcional a $\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|^{1/2}$

*Criterio discriminante*:

$$\mathbf{W} = \arg\max_{\mathbf{w}} \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_C \mathbf{W}\right|} \qquad (1)$$

La solución para las columnas de $\mathbf{W}$ son los *c*-1 autovectores generalizados:

$$\mathbf{S}_B \mathbf{w}_j = \lambda_j \mathbf{S}_C \mathbf{w}_j \qquad (2)$$

asociados a los *c*-1 autovalores generalizados mayores.

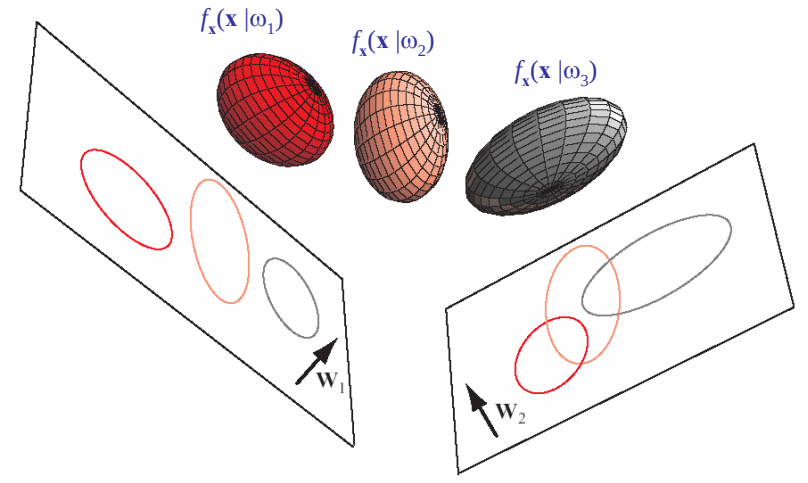


La dimensión máxima de los vectores $\mathbf{y}$ es *c*-1, ya que:

1. El rango de $\mathbf{S}_B$ es c-1
2. Por tanto solo existen c-1 autovalores distintos de cero
3. La maximización de (1) implica que solo los autovectores asociados a los autovalores distintos de cero han de incluirse en la solución.

******* Demostrad que la ecuación (2) maximiza (1) *********
******** substituyendo en (1) y usando $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ *********

Ejemplo 2:




Las distribuciones tridimensionales asociadas a tres clases distintas se proyectan sobre espacios de dimensión inferior (planos en este caso). La proyección sobre el plano definido por $\mathbf{W}_1$ proporciona mayor separación entre clases que el plano definido por $\mathbf{W}_2$.



## Regiones de decisión sobre los datos transformados

El cálculo de las regiones de decisión puede hacerse siguiendo las pautas de un detector bayesiano óptimo (ver tema 2.1) suponiendo gaussianidad para $\mathbf{y}$, con parámetros:

$$\boldsymbol{\mu}_i = \mathbf{W}^T\left(\mathbf{m}_i - \mathbf{m}\right) \qquad i = 1,\ldots,c$$

$$\mathbf{C}_i = \mathbf{W}^T\mathbf{S}_{C,i}\mathbf{W}$$

# 2.2.5 CONCLUSIONES

- Si se puede suponer una forma paramétrica para $f_{\mathbf{x}}(\mathbf{x}|\omega_i)$ entonces la fase de entrenamiento del clasificador se reduce a la estimación de los parámetros

- Pueden utilizarse dos soluciones para la estimación de parámetros: ML (más simple computacionalmente) o bayesiana (si se dispone de conocimiento a priori sobre los parámetros)

- Los problemas derivados del exceso de dimensionalidad pueden reducirse mediante el uso de múltiples discriminantes